DOCUMENT RESUME

ED 213 555 RC 013 217

AUTHOR Torres, Myriam; Arce, Carlos H.
TITLE Archiving and Disseminating Quantitative Social Research Data on Chicanos.
PUB DATE Sep 81
NOTE 32p.; Revised version of paper presented at the National Symposium on Chicano Academic Library and Information Service, American Library Association Annual Meetings (San Francisco, CA, June 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Archives; Databases; *Data Collection; *Information Dissemination; *Information Networks; Information Sources; Information Storage; Interdisciplinary Approach; Mexican Americans; *Social Science Research; Spanish Speaking
IDENTIFIERS *Chicanos; National Chicano Research Network; *Quantitative Data

ABSTRACT

An initial but comprehensive attempt to formally discuss the archiving of Chicano social science data archiving, reveals the range and quality of archivable datasets on Chicanos with a proposed taxonomic scheme for archiving them, and addresses technical and procedural issues and the strategies being adopted to establish a Chicano data archive within the National Chicano Research Network based at the Institute for Social Research of The University of Michigan. Reasons for the usefulness of a Chicano social research data archive are reviewed: economy of financial and human resources, accessibility-availability, improvement of quality of data and materials, secondary data analysis, replication, aggregation of data, and training. Four important types of Chicano surveys are identified: localized surveys using probability sampling, wide area surveys using non-probability sampling (case studies), wide area surveys using probability sampling within purposively selected primary sites (special areas), and regional or national surveys using probability selection in all stages. Three major policy issues for an archive are considered: what to archive, defining users and contact with them, costs and funding strategies. Technical procedures are reviewed: documenting, testing, cleaning, and classifying the datasets. (NEC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED213555

Archiving and Disseminating Quantitative Social Research Data on Chicanos

by

Myriam Torres
and Carlos H. Arce

Institute for Social Research
The University of Michigan

PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Myriam Torres
Carlos H. Arce

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

RC 013217

September 1981

Revised version of paper presented at the National Symposium on Chicano Academic Library and Information Service (American Library Association annual meetings). San Francisco, June, 1981.

This paper will be part of a forthcoming (1982) anthology to be titled "Chicano Library and Information Services."

Most systematic efforts by information specialists on Chicanos and other Hispanic groups in the U.S. have centered on library science and history. Most efforts have also focused exclusively on printed and graphic materials, including historical documents, published and unpublished writings about Chicanos, artistic productions, and original literary manuscripts or personal records. Considerable advances have been made in the development of efficient and accurate schemas for organizing material and in the use of computers for storage and retrieval of such information. The magnitude of Hispanic information resources that have been identified belies the prevailing notion that there is a dearth of material about Chicanos. In dramatic contrast to this overall growth of new bibliographic efforts, there has been an almost total relative neglect in the area of social science data on Chicanos. Like bibliographic information not specifically about Chicanos, general social science data have been collected, synthesized, stored and organized for efficient and economical retrieval. However, virtually no archival projects have revolved around the now extensive data on Chicanos that has been collected in numerous locations and on diverse topics through censuses, surveys, experiments and other social science data gathering methods. To date, there has not even been an attempt to conduct an inventory of such datasets.

The present paper is an initial, but comprehensive attempt to formally discuss the archiving of Chicano social science data. It first provides a context in which this effort can take place and then reviews in considerable detail the problems and strategies inherent in the effort. Wherever possible, we illustrate the problems with actual examples of Chicano data and we elaborate on the strategies by describing the data archiving activities we have undertaken. More specifically, this paper establishes the context by outlining and discussing the most significant objectives and principles regarding social science data archiving and by reviewing types of "archivable" data on Chicanos. We especially focus on our efforts to establish a data clearinghouse within the National Chicano Research Network (NCRN), based at the Institute for Social

Research (ISR) of The University of Michigan.

As suggested above, social science data archiving shares some basic concepts and objectives with other types of information archiving. Of particular importance are the goals of preservation, centralization and secondary application. However, social science data archiving differs in several important respects from other information archiving. It requires the storage of data in computers in a form which allows for exploration and analysis primarily within the computer. As such, the technical requirements for storing and structuring the information entered into such an archive are more complex than they need be for bibliographic/documentary archives.

Another distinguishing characteristic between these two types of archiving is that social science data are typically about individual subjects and a "file" or dataset will generally contain the raw information on individuals or "cases" from original surveys, evaluation projects, experiments, or other studies. Computerized literature archives, on the other hand, are generally not about individuals nor composed of raw data, but rather deal with aggregate and/or processed information. Thus, a set of articles, some published and some unpublished, based on a survey of Chicano students at a college may appear as a set of individual citations, with index terms to permit searches by subject, and optionally with an abstract to allow initial assessments of relevance or significance by a user of the bibliographic archives. The original data about these students on which the papers are based, may also be computer stored, according to certain technical specifications and conforming to human subject protection standards; such information would be stored and disseminated in a social science data archive. A specific example may be minority college enrollment patterns. One could search for relevant writings in the extensive computer based bibliographic archives which cover educational topics (e.g. ERIC), or one could search out archives of original or primary data, such as those of the National Center for Education Statistics, to find specific datasets that have the figures needed for analysis.

The distinction made above may not always be as clear as the example given. There are situations, mostly hypothetical ones, in which a social science archive could contain information about materials or documents. To illustrate, a researcher could do content analysis of a community's newspaper to explore the incidence of references to Chicanos over a period of several years and to assess the patterns of negative and positive references. The mentions and their assessment (favorable to Chicanos or not), would be raw data which could be archived and used for analysis by an investigator other than the original researcher.

This paper is organized into three major parts: 1) the principles and objectives of social science data archiving, with some reference to Chicanos; 2) the range and quality of archivable datasets on Chicanos with a proposed taxonomic scheme for archiving them; and 3) the technical and procedural issues, and the strategies being adopted by us as we establish a Chicano data archive.

Social Research Data Archiving

With the rise of technological advances in the early 1960s, there have been tremendous improvements in the development of electronic data processing equipment such as computers, mini- and macro-computers, and other equipment. The impact that these developments have had on the storage and processing of social and political research data has been enormous. Storage of data was a perennial problem prior to the availability of the magnetic tapes and disks that are now used. Computer readable punched cards was previously the most widely used method. These cards were very difficult to handle when large datasets were involved, occupied much storage space, and were easily lost, disordered, or damaged when handled. While occupying significantly much less space than cards, disks and magnetic tapes can also store very large quantities of data. Several datasets can be stored on a single reel of tape thus maximizing its use.

Technological improvements have affected several aspects of processing data.

First, accuracy has increased enormously. Prior to the development of computers, many more errors were made through reliance upon hand calculations. After computers were introduced, with cards as the main input medium, mistakes were often made when cards were handled and damaged. Secondly, the speed at which analyses can be performed has increased greatly. Computer programs written to do numerous calculations on the data have greatly reduced the many hours required to do laborious hand calculations. Magnetic tapes enable the data to be read much quicker than if the data were stored on cards and had to be passed through a card reader each time a job was run. Disks enable any segment of the data to be read without starting from the beginning of the dataset, as is the case with tapes. Batch use of a computer is now being replaced by interactive (or time-sharing) use which allow an operator to work at a terminal and interact from there with the computer. This saves waiting on the queue for the results and also allows one to correct most mistakes in a job during execution. All these improvements in speed have been extremely valuable to the social scientist since data no longer need to be limited to a case study or to a very small sample size.

Finally, the development of statistical software packages have made widely available many previously unfeasible analyses than can be performed on data. The types of analyses done no longer need be limited to simple univariate or bivariate tables with frequencies, percentages and descriptive statistics, but can now cover many complicated calculations such as correlations and time series analyses. The availability of software packages has also expanded the types of users of the data to include social researchers who need know very little about the use of computers. Data analysis no longer must be limited to a computer specialist who can write a program each time a different analysis is performed. Once the data is in its final usable form, it can be used by anyone who learns the simple instructions that each program requires. OSIRIS, the package developed by the Institute for Social Research, also provides powerful programs for data management such as preparation, correction, transformation, copying, combining, reduction, and subsetting of datasets.

All of these computer technological innovations have led to the development of social science data archives. Over a space of a few years, several archives were created to serve the research and education needs of social scientists and to advance further quantitative social research and computer software technology. The general objective of creating a data archive is to support research and training for social scientists and specialists in other fields such as education, history and public administration, who have an interest and need for working with social research data. Here we list and review some of the more important reasons why a Chicano social research data archive can be extremely useful.

Economy of financial and human resources

The cost and time of conducting surveys and original studies are usually very large. Even small local surveys, if executed with the appropriate techniques for selecting probability samples can be very costly. Budgets of several hundred thousand dollars are now commonplace for surveys covering a fairly accessible population and including 1200-1500 respondents. Some of the necessary steps involved in conducting most surveys prior to the analysis stage include: selecting the sample, preparing the questionnaire, recruiting and training interviewers, conducting the interviews, preparing coding instructions, coding and keypunching the data, copying the data onto disk or tape, and cleaning the data. Even for studies with very short questionnaires, each one of these steps require considerable personnel and financial resources. The data archive provides a means for sharing costly data that already have been collected with those who cannot or need not collect data themselves. It avoids unnecessary duplication of effort and resources if a study that serves a researcher's needs has already been done.

Accessibility-availability

It takes a researcher considerable time and effort to search for and acquire data, especially if the existing data on Chicanos are dispersed throughout the country. The procedure for identifying existing and appropriate datasets is usually through information from colleagues, or through extensive library searches. A consequence of this method, inefficiency notwithstanding, is that datasets suitable to particular needs frequently remain undiscovered, resulting in reliance upon data that are located but are less suitable. Some researchers simply opt for not analyzing data but instead use tables and results published by others. This is sometimes acceptable and adequate, but if it is the only method used for analyzing data, it places significant limitations on the researcher's findings and conclusions.

A catalog-type booklet with a listing and explanation of each dataset available in a data archive is essential for allowing the user to learn what datasets exist in the collection, choosing among several datasets and, in some cases, choosing a few variables from several datasets, as will be explained further in our discussion of aggregation. The listing of the dataset holdings in a data archive also requires a subject index where each subject or variable found in the archive would be listed in alphabetical order. For each subject/variable, the datasets containing that subject/variable would be identified. The fiscal and human resource advantages inherent in a single repository of Chicano data would thus be maximized.

Improvement of quality of data and materials

Many technical improvements to some datasets may be made after they have been placed in a data archive or even prior to their inclusion. Chicano researchers may benefit from the availability of such improvements in at least three ways. First, the overall quality of data on Chicanos will improve since one of the functions of an archive

is to "clean" datasets of general interest and improve their documentation if needed. Second, datasets that would not otherwise be accessible or even known to exist by the general public due to a lack of minimal documentation, or other reasons, may be made readily available after improvements are made by archival staff. The resources and equipment needed to undertake these tasks may not have been available to the original investigator. Third, when some datasets are technically upgraded it is possible to make them available to users in different forms of input medium or to perform transformations to change the structure of the dataset. One example where both types of services are provided is when a user requests that data be printed onto cards. If the data has a logical record length (LRCL) greater than 80, an operation needs to be performed to reduce the record length to 80, add a variable to identify the case (or interview) on each record (or card) and another variable to identify the record (or card) number. Documentation will then be needed to indicate the new location of the variables, and lastly, a set of cards needs to be punched with the newly created dataset. A program available in OSIRIS, the basic software used at Michigan, performs each one of these functions. The process of dataset evaluation, checking against documentation, cleaning, improving or upgrading is explained in more detail in the last section of this paper.

Secondary data analysis

One definition given for secondary data analysis is the "extraction of knowledge on topics other than those which were the focus of the original surveys." This kind of analysis gained popularity in the fifties when data archives began to increase in size and number. One of the original purposes for creating archives was to provide a source of data from which researchers could extract a wealth of information.

There are, of course, some limitations to secondary analysis. At times, one may not be able to find an appropriate dataset. This is particularly true in research on rare or relatively unstudied populations. Some datasets may not contain variables that are

essential for analysis. Aggregating data from several datasets, as a solution to not having enough cases, is explained in more detail in a subsequent section. Another limitation is that the coding scheme provided by the original investigator for some questions may be incomplete. The sample size may not be adequate in other cases, and the list of limitations could continue. In addition to problems inherent to datasets, there are limitations on the researcher. Before any secondary analysis can be performed, some special skills and procedures are needed. Hyman, in Secondary Analysis of Sample Surveys (1972), explains in detail the technical principles that need to be used when doing secondary analysis of survey data. He illustrates how these skills and judgements come into play by presenting cases where investigators apply them for analyzing problems. Despite all these limitations, for many situations, the advantages still outweigh the disadvantages.

As previously mentioned, lack of time, money and human resources are often factors that make secondary analysis highly appealing and absolutely necessary. It is also not unusual to find primary analysts who exhaust the funds needed to perform analysis on their data before completing that stage in their study. These investigators appreciate analysis done on their data by secondary researchers who have the funds for this purpose but not the time, money and personnel to collect original data.

Secondary analysis is also very helpful for examining historical trends. Data collected in the past serve the needs of the researchers of the present in two ways. First, they provide a way for the investigator to understand the values, attitudes and circumstances surrounding a particular time period. We could never go back in time and try to understand social attitudes without data collected during that time. The second advantage to having historical data is for comparisons with the present or between any two or more points in time. The need for this type of comparison is widely recognized in social research today. Many studies plan to re-interview in order to examine individual-level change and trends of attitudes on specific issues or to obtain some measures of social progress, or lack of it. But, at times it is not possible to incur the expense of

conducting second interviews, or it is not feasible to wait ten or fifteen years to elapse before any comparisons can be made. Secondary analysis enables one to examine these trends or changes by using data collected ten or fifteen years earlier. At the same time, researchers of the future would be able to use our current data to compare with data collected in the future.

These general benefits obtainable from secondary analysis of data are not the only ones. There are other benefits that are particular to each researcher's situation and data archives provide an excellent tool by which such analysis can be facilitated. The secondary analysis strategy is particularly important for Chicano researchers and for other researchers interested in Chicano data. Chicano academics are generally more likely to be affiliated with institutions lacking the resources and facilities to support major data gathering efforts. For such researchers, secondary analysis and local replications of large studies previously done by other organizations provide valuable and realistic opportunities for engaging in high quality research. Chicano social scientists are also more likely to have had fewer opportunities to acquire rigorous graduate and post-doctoral training in research methods. Secondary analysis on well documented and carefully archived data can develop or sharpen skills for the researcher who wants improved skills but is beyond the conventional learning phases of graduate school training. Other special, compelling reasons why Chicanos should engage in secondary analysis of archived data have more to do with external factors. Three may be cited in this context: 1) governmental and private support for social science research has seriously declined in recent years, thus making less expensive secondary analysis more attractive, as suggested earlier; 2) data on Chicanos have expanded rapidly in recent years and in periods of such expansion there are likely to be substantial underanalyses of data; it would therefore be expected that many Chicano datasets are not being fully exploited; 3) the data on Chicanos that have been collected, especially in those projects conducted before the early 1970s are more likely to have been analyzed with paradigms which Chicano social researchers have discredited or criticized; in many such cases it is

possible to reanalyze data using different theoretical or conceptual paradigms which are less biased and more credible vis-a-vis the Chicano reality.

Replication

This technique of conducting social research is a further advantage of secondary analysis. Many of the methods of replication studies involve using secondary data and require the application of principles and procedures that are used in analyzing secondary data. Instead of including replication as a type of secondary analysis, we discuss it separately because of its value as a technique and because of its current widespread use by social scientists. There are three different ways of creating a "replicated" survey. Hyman's names for these three methods are: 1) internal replication with multiple surveys, 2) external replication by the secondary analysis of a single survey, and 3) intra-survey replication-multiple indicators. These methods can be used separately for a study or can be used simultaneously within a single research project.

The first method includes using several independent surveys that were not originally intended to be linked to each other but contain identical or equivalent indicators. The data from these parts of the surveys are then linked by a secondary analyst into a single dataset which contains what can be called internal parts in a common secondary analysis.

In the second method, a replication is made of a survey, and then another replication is made of the last survey and the number of replications can continue indefinitely. An investigator can then link each replicated survey after some time. Each survey is not necessarily conducted by the same investigator. The important factor that distinguishes this method from internal replication is that the separate studies were not independent of each other but were deliberate replicates of each other. This type of replication is used widely by the Institute for Social Research at Michigan where the same or very similar interview is used in a new survey after a period of time has elapsed.

The last method involves replications within the instrument used for a single survey. Several questions measuring the same variable may be asked; these may or may not be combined at a later stage in the analysis of the survey to form indices or scales. When a researcher begins to design his/her questionnaire, secondary datasets are very valuable since it is highly unusual for a questionnaire to be totally new, without any questions or sections from previous surveys. Many standardized and tested scales such as indicators of well being, depression, mental health symptoms, stressful life events, and many more already exist. In a survey being designed, an investigator may use several established indicators just as they are, may add items to them, or may eliminate some questions. He/she may even use the published results of the survey or perform secondary analyses on the data in order to decide on the final form of the questionnaire.

Aggregation or pooling of data

Aggregation is another technique of secondary analysis of social research data similar to replication in its objectives. The goal is to increase the size of the sample when dealing with a problem of a sample size which is too small or a demographically rare population. As in replication, indicators or variables that are identical or similar in separate surveys are used, whether these surveys were independently conducted studies or not. For example, several national surveys may have a reliable ethnic origin question by which all Chicanos interviewed can be identified. If the sample size of Chicanos in each survey was approximately 50, it is virtually impossible to infer from any single survey anything about the population of Chicanos. However, if the data for the Chicano respondents from five separate surveys are combined, then the sample size can increase to 250 cases. Aggregation differs from replication in that after the surveys are pooled, the aggregated survey is then treated as if it were a single study and analysis is conducted on the pooled data as if it were a single dataset. As a result, the identity of each survey is lost. In replication, the analysis of each linked survey is generally done

separately. The inherent problem of pooling data collected at different points in time, especially with a population which may be shifting demographically and in other respects, may often be compensated by the pragmatic need to do statistically sound tabulations with a large enough sample of Chicanos. The decision whether to pool cases from a separate sample may thus be dependent on resources and feasibility, type of social phenomena being studied, and range in time of data collection of the various surveys pooled.

Training

An integral activity emerging from the maintenance of a data archive is to provide methodological and technical training. The range of persons who can benefit from such training activities includes PhD level social scientists, graduate students, and other persons whose jobs require prior training in research techniques and analysis. Some colleges and universities are not able to provide a curriculum that covers the spectrum of quantitative methods that students and other researchers need. With the rapid advancement in technologies and the new techniques and equipment constantly being developed, most institutions cannot provide their students with up-to-date training programs. In addition, many Chicano graduate students and PhD recipients, even if educated at major institutions, have not had sufficient data analysis training during the formal phase of their graduate instruction.

The National Chicano Research Network provided summer courses and workshops for Chicanos and other Hispanics between 1977 and 1980. The overall objectives of these training programs were to promote the formation of collaborative relationships among researchers and to help Chicano researchers acquire skills in quantitative social research techniques. In addition to various topical and methodological courses, there were workshops or laboratory sessions in which participants were instructed in the use of computer data accessing and processing methods. The data used for these sessions were

on Chicanos.

For most of the Chicanos participating in our training activities the material taught in the courses on elementary analysis methods and the use of the computer were totally new. This intensive training provided them in a short period of time with skills usable in their theses analyses or in their work situation. An added feature in these seminars was in being able to use data on Chicanos for practice and analysis. This provided participants with insights into the nuances associated with analyzing data on Chicanos as opposed to data on whites. Without archived data, such activities are difficult, if not impossible to conduct.

Another way in which individuals are trained through the existence of social research data archives is in the availability of clean, well-documented data for use in classroom exercises. Many graduate departments have recently begun to incorporate the use of existing datasets as a part of their instructional program. An archive provides a rich assortment of data from which a professor may choose subsets and incorporate them into the regular curriculum for teaching on research methods. The common practice in courses on analytical techniques has been to use very small datasets from the instructor's own studies or for the professor to assign values for a small number of variables in a hypothetical study. The problem with the first instructional approach is that not all professors can conduct a study themselves and if they do, the study may not contain variables that are suitable for all courses. A significant drawback with the second practice is that, in most cases, fictitious data do not contain the real problems that most datasets have. There is usually no missing data; the range and values of the variables have been perfectly assigned so that no transformations on the values are needed; the distribution is usually normal or skewed in the right direction; and so on. As a result, the real-life problems that arise when analyzing data are completely eliminated. By using survey data from an archive of real social research data, students will have more comprehensive and realistic training in analysis techniques.

A final training benefit resulting from the maintenance of a social science data

archive is the capacity for providing technical assistance and/or training to primary researchers on how to properly organize, document and clean their data in order to make them ready and available to secondary users. Guides or packages on procedures for such preparation of data can be developed as a secondary product of an archive and distributed to interested researchers. In the case of a Chicano archive, as Chicano researchers affiliated with the archive become more familiar with the steps involved in preparing data for analysis, a kind of self-feeding process would result. The quantity and quality of data in the archive would increase, resulting in even better, more widely-used techniques which would further increase the size and utility of the archive.

Archived and Archivable Chicano/Hispanic Data

Although data from several types of social research efforts could be archived, most social science archives contain primarily data from surveys. Since the NCRN Data Clearinghouse will initially consist largely of survey-type data, it is more appropriate to limit our discussion to types of surveys which contain data on Chicanos. It is imperative to note first that Chicanos are widely scattered in many parts of the U.S., even though the majority reside in ethnically segregated enclaves (barrios) and in specific regions (the Southwest and Chicago). Secondly, it should be understood that Chicanos constitute only about 5% of the total U.S. population; it is essential to recognize this statistical minority status which in survey sampling terminology is called rare element population. The principal consequences of the size and distribution of the Chicano population are that: 1) it is relatively easy to generate a representative sample of the barrio populations (e.g. those Chicanos living in conditions in which they constitute a majority-- if not nearly the entire population-- in a specific, well defined area, such as a municipality or a group of census geographical units), and 2) it is exceedingly difficult to draw a sample that is truly representative of all Chicanos because scientific and unbiased selection of those Chicanos who do not live in Chicano enclaves but still constitute a significant proportion

of the total Chicano population is like looking for needles in haystacks. Our discussion of data includes a review of several types of Chicano surveys and is followed by brief descriptions of existing social science data archives and their holdings of Chicano data.

Classes of Chicano survey data

Direct, efficient and economically feasible collection of reliable data from individuals belonging to a rare element population (such as Chicanos) has been shown to be a complicated and nearly impossible enterprise. Surveys of racial or ethnic minorities in the United States have in recent years gained in importance both in academically based research and in commercially oriented polling. By their nature, these sample surveys have confronted numerous conceptual, operational and technical difficulties which have not been adequately discussed in the social scientific literature. Surveys of Chicanos (and of other Hispanics) are now numerous and have for several years been conducted in many areas and localities by researchers focusing on diverse phenomena. Four types of Chicano surveys can profitably be discussed briefly here: 1) localized surveys using probability sampling; 2) wide area surveys using non-probability sampling ("case studies"); 3) wide area surveys using probability sampling within purposively selected primary sites ("special areas"); and 4) regional or national surveys using probability selection in all stages.

There are too many local studies which have utilized probability samples of households for us to adequately review or even list. However, two examples of such surveys which have used somewhat different methods are the Los Angeles Fertility Study and the San Antonio Study of Mexican American Elderly. The Los Angeles study sought to research fertility behavior of Chicana, black and white women. In order to draw the sample of Chicanas, the investigators generated an extract of telephone numbers listed for persons with Spanish surnames and conducted ethnic eligibility screening interviews by telephone in those households. In the San Antonio study, the researchers' objective

was to look at social services use, particularly in terms of how knowledge of help services was obtained. In order to achieve this, a probability sample of elderly persons was drawn at each sampling stage beginning with census tracts. These types of studies are often conducted with considerable rigor and are carefully documented. In spite of the non-comparability of methods in defining and selecting Chicanos, these can be welcome additions to a Chicano data archive.

Various "case study" sample surveys of Chicanos and other Hispanics have been conducted at the national or regional level. A typical example of this type of study is the Hispanic elderly study carried out between 1979 and 1980 by the Asociacion Pro-Personas Mayores. In this study, high density counties in the U.S. were selected, omitting the Midwest and Pacific Northwest regions. Five counties were purposively selected and a random sample of several counties was made out of the remaining counties. Quotas were set for maximum and minimum number of respondents. Although guided by instructions intended to minimize the selection bias and to maximize the diversity of the Hispanic elderly respondents, the interviewers had considerable latitude and could easily have contributed unknown biases into the selection process. In spite of inherent limitations, this type of study would be a useful addition to a Chicano archive if adequately documented. The adequacy of documentation and the probable difficulties inherent in these studies are mentioned in the final part of this paper.

In the category of "special area" surveys which use non-probability selection of primary sites and probability criteria in subsequent selection, there is one prominent and well known example, the Mexican American Study project conducted by UCLA in the mid 1960s by a team of investigators headed by Leo Grebler. In this study, the survey portion was done by restricting the area of study to the Los Angeles and San Antonio metropolitan areas. After this non-probability selection of the two primary field sites, the researchers sampled Mexican origin households from a carefully stratified matrix of local areas. With known probabilities of selection, dispersed Chicanos in predominantly black or white neighborhoods, as well as those in the Chicano barrios of San Antonio and

Los Angeles were sampled. This is one of the few surveys focusing specifically on Chicanos that one could expect to find in a social science data archive. However, to the best of our knowledge, it is not available in any established archive. A second example of a large scale survey project is the forthcoming Hispanic Health and Nutrition Examination Survey (Hispanic HANES). As proposed, this survey would utilize precise methods of selecting households, but only after the study areas are purposively chosen. When completed, the data from this survey will certainly be included in several archives.

There is only one national probability sample survey which generates large samples of Chicanos and other Hispanics. This is the regularly conducted Current Population Survey of the U.S. Bureau of the Census. Every March, and at selected other times when special surveys are conducted, the Current Population Report publishes data on Hispanics from their surveys. By rotating the monthly sample, in which individual households are in the sample for four months then out for four months and then in the sample again for four more months, the CPS is able to "accumulate" data for four months of distinct respondents to achieve the "large" Hispanic sample every March. The CPS uses primary sites selected from the 1970 census and updated address lists are used to select households. Data from these and other surveys conducted by the Census Bureau are available in many archives. Because these archived data files are often of prohibitively large size and complexity, and have only small percentages of Chicano respondents, it is frequently necessary to create "extracts," comprised of the Chicano cases, with or without a sample of other cases. Such special extracts could also have a simpler structure and be limited to relevant variables, thus making their use more economical and technically easier. These extracts would form an important part of the NCRN Data Clearinghouse.

In the context of the variety of surveys that have been conducted on Chicanos (and it should be noted that we cite only a very few of the large number of studies that could have been given as examples), the value of a Chicano data archive becomes apparent. Only through centralized and uniform storage and dissemination of diverse datasets can

researchers hope to maximize the data that have been collected. Chicanos have indeed been studied but with methods that are diverse and of mixed quality. The archive allows for use of multiple sources of data and permits researchers to extract the most useful of each source, in the form most adaptable to their needs.

Identification and availability of Chicano data in existing archives

Although most established social science data archives have data on Chicanos or on an aggregated "Hispanic" minority, none have made concerted efforts to strengthen, expand or adequately organize their empirical data on Chicanos. In this section, three major existing data archives are described.

The Inter-university Consortium for Political and Social Research (ICPSR) is located in the Institute for Social Research at The University of Michigan. It was established in 1962 by political scientists from twenty-one universities primarily to serve political analysts. It now provides data pertinent not only to all social sciences but also to fields of education, public health, social work, public administration and many others. The Roper Public Opinion Research Center located in Williamstown, Massachusetts, is a national and international resource data bank containing over 13,000 datasets (dating from 1936) contributed by 122 survey research organizations in 74 countries. It was established in 1946 to house Roper's own surveys and in 1957 it was officially organized as a general archive. Currently, it is the largest and longest-established repository of survey data. The National Opinion Research Center (NORC) data archive contains a general collection of the Center's own surveys of the United States' national population and of special areas within the country. The Center's study directors use survey methods to conduct extensive research on basic problems in social sciences and to study specific problems in such areas as education, religion, social change, ethnicity, political participation and leadership, as well as survey methodology. The maintenance of an

archive is a secondary function of the Center while its primary function is the conducting of surveys.

Each one of these archives contains very few or no datasets that are exclusively on Chicanos. Many datasets, however, especially those generated by Census Bureau surveys, do contain some data on Chicanos. Others can be safely assumed to contain such data, but the indexing procedures and dataset descriptions for these do not permit a potential user to ascertain the relevance of a particular dataset for an analysis of Chicano social conditions.

The NCRN Data Clearinghouse

In this final section, we review the problems encountered in the actual creation of a Chicano social research data archive, and the solutions we have generated. This project is a component of the National Chicano Research Network, created to serve various needs of Chicano social researchers. The current primary activities of the Network are training and information sharing. As a natural extension of these activities the Network has established a data clearinghouse which seeks to meet the general objectives of social science data archives but with a special focus on Chicano data. We discuss the effort in three areas: policy issues, technical procedures and provision of services.

Policy issues for an archive

At least three major policy issues need careful review before fully operating a data archive: criteria for deciding what to archive, defining the constituent/user population, and cost and funding considerations.

What to archive. It is neither feasible nor necessary to archive every dataset that is exclusively on Chicanos or that contains a sample of Chicanos. Initially, the most practical approach is to process those datasets already in our possession and familiar to

us. In addition, data should not require extensive cleaning and should already have proper documentation. After these initial datasets have been processed and archived, an on-going selection and acquisition process must be established. The National Chicano Research Network (NCRN) Coordinating Committee makes decisions as to which datasets will be acquired for the archive. This committee is composed of six Chicano social researchers. At each bi-annual meeting of the committee, a list of possible acquisitions is reviewed and prioritized for purchase. Depending upon available resources and demand, it is anticipated that all datasets that receive a reasonable priority rating would be purchased eventually.

The NCRN Data Clearinghouse bases its acquisition selections on several factors. Since the Chicano archive will be affiliated with the ICPSR archive, the committee will apply many of the same considerations for acquisition of datasets that ICPSR currently uses. Some of these include: 1) whether the data still exist and are computer-readable; 2) if the data collector is willing to have the data archived; 3) cost to acquire the data; 4) current and future research value; 5) whether the data complement a collection already archived or represent a desired new data area for the archive; 6) if the collection represents a major research contribution in a given discipline or sub-discipline; 7) time span of the collection; 8) instructional value of the collection; 9) whether the collection is longitudinal (collected at several points in time) or might be replicated in the future; 10) whether the collection is part of a continuing series such as census materials or election returns; 11) whether the data are tied to specific software systems since data that are software-dependent will be more difficult and costly to handle; 12) if adequate documentation for the data is available; and 13) if confidentiality problems exist.

The weight given to these considerations may vary, depending upon the topics covered by the particular dataset and how rare they are. For a Chicano data archive, it is particularly important to note that because of a dearth of empirical data on a particular topic, it is necessary to be flexible on the technical criteria when the data in question are important for their uniqueness. For example, a dataset from a survey of

unapprehended undocumented residents may have serious problems of technical documentation but may be so uncommon that it should be archived in spite of these technical difficulties.

Definition of users and contact with them. Two characteristics typify the users of an established social science archive. First, users generally have been, or have known, prior users of the data archive, thus suggesting some familiarity with procedures and with the capacity of the data archive. Second, users typically have access to an infrastructure or network of experts. The combination of familiarity and infrastructure enhances the quality and efficiency of data archive services and products.

The individuals and organizations likely to become users of a Chicano Data Clearinghouse will generally be very different from the typical user/consumer of an established archive's products and services. In the first place, the Chicano archive clients may be much more diverse. Although some users will be researchers with considerable experience in data analysis, it is likely that many prospective users with minimal experience will be attracted to the resources of this type of archive. Among these "new" clients, there are likely to be some with modest training or experience and/or who are at very early stages of their academic careers, e.g., still in graduate school. Further, other clients may likely be individuals without research training or interest; they may represent organizations or agencies interested in empirical descriptions of a particular population or problem. Thus, their interest in the archive may not actually focus on the raw data in the archive, but on completed analyses performed on some data according to their policy-oriented specifications.

The implications of this "new" user clientele is that innovative services and communications mechanisms need to be developed. Services are discussed in a subsequent section. The latter, contact/communication with users, is particularly important for at least two reasons. First, the products and services of an archive need to be publicized and promoted among the projected clientele. It is important that regularly

published materials be widely distributed and that any catalogs of archive holdings be updated and disseminated in a standardized and systematic fashion. The second important function of communication mechanisms is to maximize clarity and accuracy regarding the technical specifications of data and computer materials, such as tapes, which have to be prepared for users. Thus, it is critical that specification forms for describing and ordering products be precise and fully detailed so as to minimize errors which can be expensive and time consuming.

Costs of archiving and funding strategies. A social science data archive can be a costly enterprise. High costs can be incurred in wages and fringe benefits for technical and administrative staff who process acquisitions and orders, render services and prepare data for archiving and dissemination. Acquisitions are frequently purchased, although at times they can be secured through donations or exchanges. Materials and equipment of many kinds, especially for doing computer work, are expensive and essential to an archiving operation. The costs of computer time and software development or purchase are additional items of expense. Finally, indirect costs or overhead, particularly for a facility that is especially designed and operated for social research, is considerably high.

Most established archives are supported by universities and other organizations that maintain membership and pay annual fees in exchange for products and services; or by federal government and private agencies that provide grants and contracts to institutions that maintain data archives. In the case of a newly-created Chicano archive, there are notable difficulties in establishing a financial operational base since few institutions would have sufficient interest to justify membership fees at a level which would generate the needed revenue for the initial creation and maintenance of a data archive. Full support from grants may be feasible in the early stages of a data archive's operation but would be difficult to maintain on a more permanent and long-term basis.

The most practical solution is a modified "fee for actual services rendered" strategy. Ideally, each user would pay a portion of the cost for each product or service

provided. With the intent of making the cost to the user more reasonable and attractive, an archive could justify and secure grant funding which would provide a subsidy. This is clearly easier to achieve if the Chicano archive is located in an established organization in which technical advice is readily and economically available, thus increasing its prospects for securing funds from both governmental and private sources. In the actual case of the NCRN Data Clearinghouse, the archive's administrative functions are handled within the overall activities of the Network and the staff and product-acquisition costs are supported by two private foundations. After a two year start-up phase, it is anticipated that costs will be modest and fundable by sponsors. Finally, as suggested earlier, the self-feeding phenomenon that can be expected in an effort such as this should attract, at no cost, new acquisitions in excellent technical condition that would not require excessive and expensive handling.

Technical Procedures

NCRN collaboration with the ICPSR data archive requires a joint effort between both organizations in the process of technically preparing a new dataset in the archive. The NCRN Data Clearinghouse is making use of ICPSR's experience in archiving thousands of datasets. Standard procedures and computer programs have been developed in order to process each dataset efficiently. Several steps are involved in preparing each dataset from the moment it is received to when it is assigned a permanent place in the archive and thus becomes ready for dissemination. It is valuable to identify and review several important phases in the preparation of datasets.

Documentation and testing of the dataset. The length of time involved in executing this step for a particular dataset depends largely on its condition when received. The process can be long and difficult for data that has poor or insufficient documentation and/or is stored in a tape that cannot be read or matched properly with

the documentation received. The documentation needs to be clear enough for reading and needs to be complete. Completeness means that the location and description of each and every variable needs to be described and fully accounted for. At times, there are additional explanations required for those variables coded in a particular manner. Other information that would be helpful to the user (such as sampling and field procedures used, or related publications) needs to be but often is not included with the documentation. When any such details are missing or discrepancies between the data and the documentation arise, the principal investigator must be contacted and the costly, laborious task of reconciling discrepancies and filling in gaps undertaken.

At the same time that the documentation and data are being matched, the actual data are checked to see if they can be read by the computer at the archive's installation. Most of the datasets are received on magnetic tapes and often the investigator will not send sufficient technical information about how the data were copied onto the tape. This step of checking the data and documentation is performed principally by NCRN's technical staff with advice from experienced archiving experts.

Cleaning and documenting. A further service provided by the NCRN is to clean a dataset when incorrectly coded values or inconsistencies are found. Some principal investigators do not have the resources to perform all the necessary data cleaning and simply analyze the data immediately after coding. Cleaning by NCRN staff can be done in one of two ways. In some instances, the data can be cleaned by simply assigning missing data codes to all those variables containing values outside the range of valid codes and to variables that are inconsistent with other variable values. If the resources are available and if a certain dataset is very valuable for Chicano research, it is feasible and appropriate for the NCRN to incur the added cost of cleaning data by referring to the instruments themselves to determine the correct values of variables rather than losing them to a missing data value.

The same type of service is provided when the documentation of a dataset is

inadequate. For some datasets that do not meet the "minimal documentation" requirements, this would be an absolutely necessary service if they are to form a part of the archive. The tasks involved can vary greatly depending on the number of variables the dataset contains and the condition of the documentation. For this reason, careful consideration must be given to the cost and effort of providing this service when deciding whether to include a dataset in the Chicano archive. If the principal investigator cannot pay for part of this cost, it is possible to engage in a service exchange as a form of payment. If this exchange is not possible, and the NCRN Coordinating Committee deems that dataset valuable for Chicano research, a maximum investment cost to improve the quality of the data and documentation is set. These same considerations and services will also be given to datasets that are already a part of the archive in order to upgrade them (see the next section on classification of datasets). As previously discussed, the overall quality, availability and accessibility of data on Chicanos will improve as a result of the technical procedures performed by the Chicano archive.

Classification of datasets. The NCRN Data Clearinghouse has adopted the same classification scheme that the ICPSR uses for its data archive. The four classes in the scheme reflect the level of technical processing which has been devoted to a dataset and not the substantive characteristics of the data. Processing goes beyond receiving, duplicating and disseminating data; it involves investing time and money in checking the data, correcting it, improving its documentation and giving it an appropriate format. In order to maintain consistency with the ICPSR classification system, the same criteria will be imposed to determine in which class a Chicano dataset belongs after evaluation of the processing done by NCRN's technical staff. A brief technical description of the four classes of data follows.

Class I datasets have been checked and corrected, if necessary, and technically formatted to ICPSR specifications. Also, the data may have been recoded and reorganized in consultation with the investigator to maximize their utility and

accessibility. A codebook, often capable of being read by computer, is available. This codebook fully documents the data and may include descriptive statistics such as frequencies or means. All Class I studies are available on magnetic tape in either card-image or OSIRIS format.

Class II datasets have been checked and formatted to ICPSR standards. All non-numeric codes have been removed. The studies in this class are available on magnetic tape in either OSIRIS or card-image format. The documentation exists as either a machine-readable codebook (which may be edited and updated as required by further processing), a multilith draft version or a photocopy of the investigator's codebook. Any peculiarities in the data are indicated in the documentation. A copy of the documentation is supplied when the data are requested.

Class III datasets have been checked for the appropriate number of cards per case and accurate data locations as specified by the investigator's codebook. Frequency checks on these data have often been made; and known data discrepancies and related problems are communicated to the user at the time the data are requested. A copy of the codebook for these data is also supplied with the data. The data themselves usually exist only in card-image form.

The Class IV datasets are distributed in the form received by the archive from the original investigator. Users of Class IV data should keep several considerations in mind. Problems may exist which would not be known before processing begins, and thus the data archive takes no responsibility for the technical condition of the data. The requestor, therefore, must be prepared to accept some uncertainty as to the condition of the data. Requests for these studies normally require a longer time to complete than more fully processed studies. In addition, technical assistance on problems encountered in the use of Class IV data is by necessity very restricted. The documentation for Class IV studies is reproduced from the material originally received and a copy is supplied with the data. The majority of the studies in Class IV are available on magnetic tape, but a few studies in this category are available only on cards because they contain multiple

punches.

In the NCRN Data Clearinghouse presently being established, all of these cleaning, documenting and classifying efforts are reflected in the technical and substantive descriptions written for each dataset and incorporated into the catalog.

Services

The NCRN Data Clearinghouse will provide services similar to those provided by other social science data archives. These services, such as copying and distributing datasets, providing advice on cleaning and documenting data, etc. have been discussed earlier. In addition, the NCRN Data Clearinghouse anticipates that services beyond dissemination or distribution of datasets will need to be provided. In ways which parallel the expected diversity in potential clientele as we discussed earlier, it is expected that the data clearinghouse will on occasion need to perform some services which are not generally provided by established data archives. The following are four significant categories of services "beyond dissemination" which are anticipated thus far: 1) provision of "custom-made" subsets of data. This may include creating tapes of a subset of variables and/or cases from an existing dataset and may occur regularly for datasets which are either very large or complex or in which the user has limited resources for managing or analyzing data. It is also possible that some datasets will be released to individual users with restrictions that allow analysis of only certain specified variables. The latter may occur with the Chicano Survey data collected by the Chicano Program of the Institute for Social Research and appropriate agreements can be made with users while the dataset is on a "restricted use" status. 2) Provision of "data" on hard copy. Most existing data archives will generally not provide printed data from a dataset. However, we expect to receive and respond to some requests for descriptive statistics (usually simple cross tabulations) from individuals or organizations which do not have data analysis resources. Such requests, if they are manageable in size and scope, will

generally be filled promptly by NCRN staff. They may then be subsequently mentioned in reports to our sponsors to demonstrate the diverse range of products and services provided. When similar requests call for more fully analyzed data requiring the expenditure of funds and resources beyond those allocated, the data clearinghouse will generally execute a contract for services which will be handled through the standard channels of our home institution and will involve regular budget requirements. The subsidized character of the clearinghouse will make such contracted services very affordable to potential users. Responses to such requests will depend on the interests, expertise and time availability of the NCRN staff. In cases where NCRN staff cannot respond to such a request, the requestor will be referred to other researchers involved in the NCRN who are knowledgeable about the particular dataset. 3) Provision of technical consultation. This service is likely to include some overlap with the two services just mentioned and to involve requests for advice or assistance in "setting up" a dataset, or performing data management or analysis at the user's installation. Again, depending on availability, expertise and interest of NCRN staff and consultants, these requests are filled on a case-by-case basis and generally require the execution of a contract for services between the requestor and the NCRN. In the last two categories of services it is anticipated that on occasion the request will be for such minor consultation that no formal project will need to be established. 4) Other consultation services requests: it is likely that NCRN will receive requests for services to research projects ranging from conducting an entire study to advising on a single procedure in an ongoing research project. Depending on the magnitude, these requests will generally be treated as "Requests for Proposals" and will be handled with a proposal and budget if the NCRN staff is willing and able to handle them.

Summary

There are many compelling reasons why it is appropriate and essential that a Chicano social research data archive be established and in operation. First, there are many datasets on Chicanos, but they are highly dispersed, are at many levels in technical condition, and tend not to have been used to their full potential. Second, existing data archives have not and are not likely to specifically attend to Chicano data needs. Third, social science writing on Chicanos has reached a stage where more empirically-based exploration of the Chicano condition is needed in order to assess developing paradigms and to more precisely document the current condition of Chicanos. We have indicated that the use of diverse datasets from an archive is the most efficient approach for this activity. Fourth, under current fiscal and labor market conditions in the academic world, it is essential that cost-efficient strategies be adopted to conduct the work of scholars. Not only is research of a secondary analysis type made possible by data archives, but such activity, with its collaborative and training opportunities, is made almost natural by a data archive.

The NCRN Data Clearinghouse is being established on precisely these principles. It is fully anticipated that, after an initial short period of operation, the Data Clearinghouse will become a two-way mechanism in which Chicano researchers and others engaged in the investigation of the Chicano condition will contribute products (datasets, ideas) to the extent that they use the data already archived. The full potential of this new data archive is difficult to estimate but is clearly very great. It requires a collaborative effort that transcends discipline, institutional affiliation or geographical location, and conceptual/ideological preference. To that extent, it closely parallels the impressive and significant efforts of the past several years in the area of Chicano bibliography, indexing and information storage and retrieval. Finally, through technical and substantive exchanges such as those made possible by this conference, the bibliographic and data archiving efforts can accelerate their development and more adequately meet the needs of Chicano scholarship.

References

Erik W. Austin & Carolyn L. Geda, "An Archive for the Social Sciences," Feb. 7-10, 1979, ISR publication # 4146.

Herbert H. Hyman, Secondary Analysis of Sample Surveys, 1972.

Research Centers Directory, 6th Edition, Archie M. Palmer, editor.

Encyclopedia of Associations, Vol. I, National Organizations of the U.S., 13th edition, Nancy Yakes and Denise Akey, editors, 1979.

Roper Public Opinion Research Center, "Survey Data for Trend Analysis: An Index t epeated Question in U.S. National Surveys held by the RPORC," 1974.

Judith Rattenburg & Paula Pelletier, Data Processing in the Social Sciences with OSIRIS. Institute for Social Research, 1974.